DOCUMENT RESUME

ED 270 492 TM 860 376

TITLE 1985 Interpretive Panel Results: Educators Set Performance Expectations. Assessment Report 13: An Update on the Alaska Statewide Testing Program.
INSTITUTION Alaska State Dept. of Education, Juneau. Office of Evaluation, Assessment and Research.
PUB DATE Mar 86
NOTE 10p.; For related documents, see TM 860 374-375.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Gains; *Achievement Tests; Cutting Scores; *Educational Assessment; Educational Trends; Elementary Education; Grade 4; Grade 8; Item Analysis; *Mathematics Achievement; Norm Referenced Tests; *Reading Achievement; Scores; *State Programs; Testing Programs; *Test Interpretation; Test Results
IDENTIFIERS Alaska; *Alaska Statewide Assessment Program; Test Score Decline

ABSTRACT

In 1985, reading and mathematics achievement tests were administered to nearly 15,000 fourth and eighth graders in Alaska. Educators comprising the Interpretive Panel developed a minimum and a desired level of performance. Scores below the minimum indicated a weakness; scores above it, a strength, and scores between, satisfactory achievement. The Panel cautioned: (1) that the tests were not minimum competency tests and that students were not expected to answer all items correctly; (2) that each school district had an individual curriculum; and (3) that reviewers should be cautious about making generalizations. Results indicated fourth graders performed best in mathematics and had trouble with every reading subtest. This was the opposite of the 1983 results. Eighth graders also showed greater strengths in mathematics than in reading. A number of skills were listed which consistently needed improvement in all three previous years of testing. An article explaining test score decline is also included. Recently, it has been found that almost every school has been using a different set of achievement tests, and being compared to a different norm group. This is because in the past year or two, developers of the SRA achievement tests, Metropolitan Achievement Tests, Comprehensive Tests of Basic Skills, California Achievement Tests, and Iowa Tests of Basic Skills have all developed new tests with new norms. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270492

1985 Interpretive Panel Results: Educators Set Performance Expectations

Assessment Report 13: An Update on the Alaska Statewide Testing Program

Office of Evaluation, Assessment & Research
Alaska Department of Education
March 1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A.E. Hazelton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

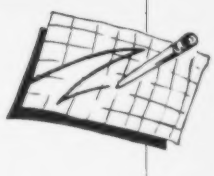
☐ This document has been reproduced as received from the person or organization originating it.

X Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 376

ASSESSMENT REPORT 13: An Update on the Alaska Statewide Testing Program



1985 Interpretive Panel Results: Educators Set Performance Expectations

In March 1985, nearly 15,000 fourth and eighth grade students participated in Alaska's fifth biennial statewide assessment. The tests measure students' achievement in those reading and math skills that Alaska educators have identified as most important. Results of the March assessment were released to schools immediately after testing. In addition, preliminary statewide averages were published in *Assessment Report 11*. We know from these results that fourth graders could correctly answer *72%* of the math items and *68%* of the reading test items. We also know that eighth graders could answer *71%* of the math items and *74%* of the reading items. But these results give us only part of the picture we need to evaluate students' performance. What we still must answer is this: Should we consider this level of performance acceptable? More than acceptable? Or unacceptable?

These are complex questions, but the state's Interpretive Panels can offer some guidelines to assist us.

The Interpretive Panel Process

The Interpretive Panel is composed of educators from throughout Alaska who were nominated by the NEA-Alaska Board of Directors in recognition of their expertise and knowledge of educational expectations. Each panel member reviewed one of the four tests (say, fourth grade reading or eighth grade math) and set two levels of expectation:

- First, a MINIMUM level: that is, the lowest percentage of students answering correctly that could be regarded as acceptable.
- Second, a DESIRED level: that is, the highest percentage of students that could realistically be expected to answer the item correctly.

Panel members worked independently, and without seeing actual test results. When they were finished, their individual minimum and desired estimates were averaged to produce a final figure against which actual student performance was judged. If actual performance fell below the minimum, it was said to indicate a WEAKNESS. When performance exceeded the desired level, it was said to indicate a STRENGTH. Performance between the minimum and desired levels was regarded as SATISFACTORY. In making their judgments, panelists were asked to consider

- The difficulty of each item relative to other items measuring the same skill,
- The importance of the skill being measured, and
- The appropriate expectations for students at the given grade level.

Given these considerations, the range of satisfactory performance differs considerably from item to item. Thus, a performance score of, say, 70% might indicate a strength on one item, but a weakness on another.

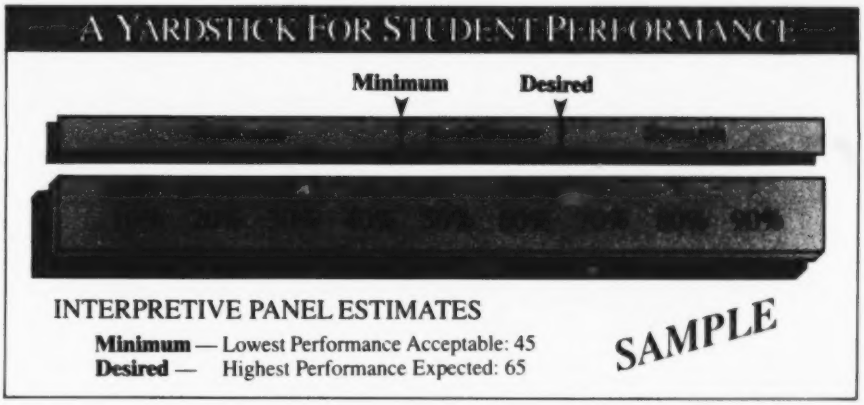
Note: See pages 5-7 for important information on potential drops in standardized test scores.

This report highlights some of the more notable findings about the relative strengths and weaknesses in each content area tested. Additional information on the strengths and weaknesses shown by the 1985 Statewide Assessment results—as well as previous years' assessments—is available in a report soon to be released by the Department of Education. A copy of this report will be sent to the curriculum coordinator at each district within the next few weeks. A limited number of additional copies are available; please contact Al Hazelton at the Department of Education, (907) 465-2966.

Evaluate the Interpretations with Caution

Once the job of the Interpretive Panel is completed, individual districts must determine how best to use assessment results. It is not within the scope of this report to offer recommendations on using results, or even to suggest what impact—if any—these results should have on state or district curriculum. But we can present some considerations to help you decide what results might mean.

(continued, next page)



Office of Evaluation, Research and Assessment, March 1986

BEST COPY AVAILABLE

Evaluate with Caution

(continued from page 1)

First, remember that the Alaska assessment tests are NOT minimum competency tests. The items vary widely in difficulty, and it would be unreasonable to expect all students to answer a given item correctly.

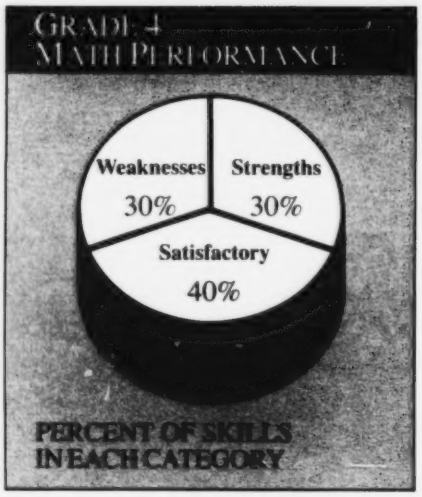
Second, Alaska has no statewide curriculum; rather, each of the state's 55 districts determines independently what content will be taught at a grade level—and that content may or may not match what's covered on the assessment tests. Though Interpretive Panel members are sensitive to these differences, apparent strengths or weaknesses should still be carefully reviewed by local educators, who are most familiar with district curriculum and goals.

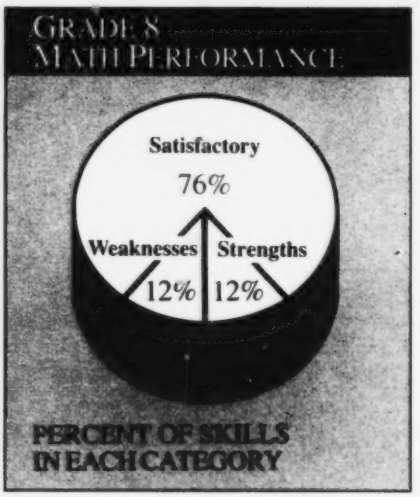
Third, reviewers should be cautious about generalizing. For example, results show that 90% of the eighth graders tested could correctly interpret a figurative expression, but that only 50% could successfully identify the mood of a poem. Had another figurative expression or poem been used, results might have been quite different. Thus, we shouldn't conclude that eighth graders will always have an easy time with figurative language generally, or that interpreting poetry will always prove difficult.

Finally, the Content Panel members who helped develop and review these tests selected items they felt were suitable in content and difficulty for statewide testing. Obviously, individual districts' needs and goals will differ, and local educators must use their own judgment in evaluating results. Still, the panelists believe that the skills they selected are important; that is, there should not be much of a discrepancy between the test content and what is actually taught. Local educators may want to examine the assessment tests to see where differences exist. Review copies of the test can be obtained for this purpose at no cost from the Alaska Department of Education. Telephone the Office of Evaluation, Assessment and Research at (907) 465-2966.

The Students' Performance: A Brief Summary

The Alaska assessment tests actually comprise six subtests at each grade—three in math and three in reading. The math subtests are Application, Computation and Concepts. The reading subtests are Comprehension, Study Skills and Word Skills.





In contrast to 1983, when fourth graders gave their best performance in reading comprehension, fourth graders in 1985 performed best on math concepts, with seven strengths, three satisfactory performances, and five weaknesses. Fourth graders demonstrated at least some strengths on all math subtests, in fact. They had trouble, however, with every reading subtest; in fact, there were no identified strengths for fourth graders in reading skills. They had the least difficulty with reading comprehension, where performance was rated satisfactory on all but one of the 15 items. On both study skills and word skills, however, weaknesses outnumbered satisfactory performances 8 to 7.

Eighth graders also performed slightly better on math skills than on reading skills—again, a contrast to the 1983 assessment, in which eighth graders performed slightly better in reading than in math. In 1985, eighth graders showed three strengths in math applications, and one each in math computation and math concepts. Most eighth grade math performance was rated satisfactory, with only five identified weaknesses over all three subtests. Similarly, most eighth grade reading performance was rated satisfactory; there were, however, two areas of strength and five weaknesses across the three reading subtests.

Now let's take a closer look at how the Interpretive Panel judged student performance for 1985—by grade level and by subtest.

Mathematics Application

Application items test problem solving skills: identifying facts, translating written information into math equations, choosing the correct approach to solving a problem, determing which data are needed to solve a problem, and so on. Items generally consist of a short story problem with one or two relevant questions.

GRADE 4. Fourth graders did best on determining the required operation (addition, subtraction, etc.) to solve a word problem, and on determining what information (e.g., total number of people, amount of money) was being asked for by a word problem. These two skills had been areas of strength in previous assessments as well. By contrast, fourth graders had difficulty determining which numbers to use in solving a word problem and in determining whether a problem contained enough information to be solved when that problem contained extraneous information. The first had been an area of weakness on the two previous assessments as well; the second had been an area of strength in 1981. **The scoreboard: 2 strengths, 3 weaknesses, 5 satisfactory.**

GRADE 8. Eighth graders gave their strongest performance in determining the information asked for by a word problem, identifying the appropriate label for the answer to a volume problem, and determining the information needed to solve a word problem. They had difficulty determining whether additional information was needed to solve a word problem and whether an exact fact was stated in a word problem. On the 1981 assessment, eighth graders had shown no strengths in this

BEST COPY AVAILABLE

area; on the 1983 assessment, they had shown only one—on determining how to solve a word problem. In addition, their 1985 performance was judged satisfactory on several skill areas that had been considered weaknesses in 1983—including finding an average. **The scoreboard: 3 strengths, 2 weaknesses, 11 satisfactory.**

Math Computation

The computation subtest measured basic arithmetic skills: adding, subtracting, multiplying and dividing whole numbers (positive and negative), fractions, and decimals. Though problems were not intended to be tricky, they were constructed in such a way that common errors would lead students to select an incorrect option. Such a format helps reviewers to make an educated guess about the cause of student errors, and perhaps take steps to improve performance.

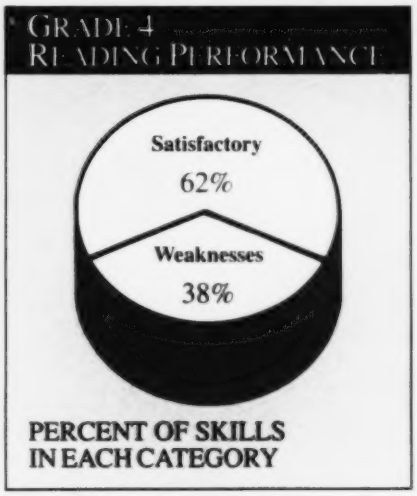
GRADE 4. Fourth graders' performance was rated satisfactory on the majority of the 12 computation items. They had difficulty with only three: all required multiplying or dividing two digit numbers. They gave their strongest performance (10 percentage points over the desired level of performance) in subtracting numbers with decimals aligned—and also showed strength in adding numbers with decimals aligned. **The scoreboard: 2 strengths, 3 weaknesses, 7 satisfactory.**

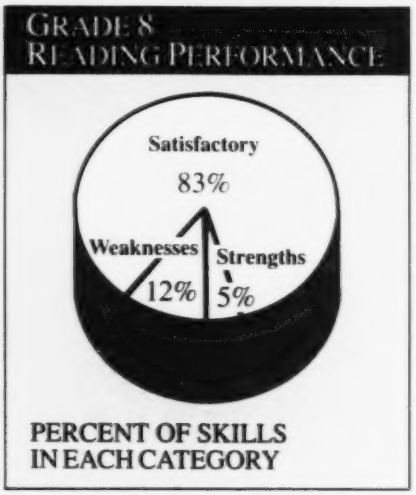
GRADE 8. Eighth graders' performance was rated satisfactory on 10 of their 12 computation items. They did, however, fall well below the minimum expected performance level on subtracting negative from positive numbers. On the other hand, they showed a strength in subtracting mixed numbers; this was the first identified strength on this subtest in all three assessment years. **The scoreboard: 1 strength, 1 weakness, 10 satisfactory.**

Math Concepts

Math concepts cover such things as number properties, place value, rounding, telling time, using units of measurement appropriately, equivalent fractions, and geometric shapes. These items measure a student's understanding of information presented, but do not require any calculation.

GRADE 4. Fourth graders gave their best math performance on this subtest, with identified strengths on 7 of 15 items. Their biggest problems were in identifying the relationship between two metric units; on both items measuring this skill they fell far below the minimum acceptable performance level. Areas of strength included recognizing fractions, rounding numbers to the nearest ten, and matching time on the clock with a written expression. On previous assessments, performance had been rated primarily satisfactory on this subtest, with only three identified strengths in 1983, and two in 1981. **The scoreboard: 7 strengths, 5 weaknesses, 3 satisfactory.**





Their biggest problems were in identifying the relationship between two metric units; on both items, they fell far below the minimum acceptable performance level.

GRADE 8. Of 15 items on this subtest, eighth graders gave satisfactory performance on 12. There was only one identified area of strength: determining the greatest common factor for a set of numbers. This skill had been identified as a weakness on the '81 assessment. Eighth graders had the most difficulty with items involving terms like parallel, perpendicular, point, line, plane, and so on; and with an item questioning the equality of an equation: e.g., $(2+A) \times 4 = (2\times4) + (A\times4)$. Equation items had given eighth graders difficulty on previous assessments as well. **The scoreboard: 1 strength, 2 weaknesses, 12 satisfactory.**

Reading Comprehension

Reading comprehension items consist of a reading passage followed by three questions based on that passage. The questions are designed to measure students' ability to identify the main idea, distinguish fact from opinion, draw inferences, reach conclusions, determine the order in which major events occurred, understand how ideas are organized, and so on. Great care is taken to ensure that the reading level of each passage is appropriate for the grade level being tested.

GRADE 4. While there were no identified strengths for fourth graders in this area, their performance was rated satisfactory on 14 out of 15 items. The only identified weakness involved recognizing editorializing in an author's statements; performance on items measuring this skill had been rated satisfactory in previous years. **The scoreboard: no strengths, 1 weakness, 14 satisfactory.**

GRADE 8. Eighth graders' performance was rated satisfactory on 13 out of 15 items, with one area of weakness and one strength. Students at this grade level had some difficulty identifying the mood of a poem; on the other hand, they showed skill in interpreting figurative language within a passage. For 1985, their performance was rated satisfactory on an item requiring them to determine whether satisfactory evidence had been presented to support the writer's conclusion; on previous assessments, performance in this area had been far below the expected minimum. **The scoreboard: 1 strength, 1 weakness, 13 satisfactory.**

(continued, next page)

BEST COPY AVAILABLE

Performance Summary

(continued from page 3)

Study Skills

Can students use a dictionary effectively? How about a map, index, graph, chart, or table? This section of the test measures students' ability to find and selectively use information they need. Readers should take special care in interpreting results from this section, since the Content Panel members who selected these items for the test were concerned that the required skills may not be part of some Alaska districts' curriculum.

GRADE 4. Fourth graders' performance was rated satisfactory on 7 out of 15 items in this area. There were eight identified weaknesses. Students had the most difficulty on items requiring them to find a topic in an index, use guide letters to locate a topic within an encyclopedia, study a line graph to answer a question, and follow written directions to answer a question. This performance was not very different from that in previous assessment years; only in 1983 was there an identified strength for fourth graders on this subtest. Readers are reminded that items within this area often require very specific skills that may not be part of every district's curriculum, or at least not covered by the time fourth graders take the test. **The scoreboard: no strengths, 8 weaknesses, 7 satisfactory.**

GRADE 8. Eighth graders' performance was rated satisfactory on 10 out of 14 items in this area. They had the most difficulty with an item asking which of several publications—based on title and publication date—would most likely contain out-of-date information. This skill had been an identified area of weakness on both previous assessments as well. Students also had some difficulty using dictionary definitions to answer a question. On the other hand, performance revealed a strength in using a map key (or legend) to answer a question. **The scoreboard: 1 strength, 3 weaknesses, 10 satisfactory.**

Word Skills

This section of the test covers skills many Alaska educators consider vital to reading proficiency: knowledge of regular and irregular vowel sounds, consonant combinations, possessives, prefixes and suffixes, synonyms and antonyms, and multiple meaning words.

GRADE 4. This was a difficult subtest for many fourth graders. Their performance was rated satisfactory on just 7 out of 15 items, and there were 8 identified weaknesses; there were no strengths. On two items—recognizing words with the same initial blends or digraphs—performance fell more than 20 percentage points below the acceptable minimum set by the Interpretive Panel. On the other hand, two areas of satisfactory performance for '85 had been areas of weakness on previous assessments. These were recognizing words with the same short vowel sound and recognizing identical vowel sounds between words containing some silent vowels. **The scoreboard: no strengths, 8 weaknesses, 7 satisfactory.**

As stressed earlier in this report, there is no statewide curriculum and it is perfectly acceptable for some of the skills not to be covered by a district. But the decision not to teach a skill should be a carefully considered decision, not an accident.

GRADE 8. Eighth graders' performance was rated satisfactory on 12 out of 13 items. They had only one identified weakness: using the correct plural possessive noun (e.g., girls' clothing) to complete a sentence. Performance in previous years had revealed some strengths for eighth graders in this area—but also more weaknesses. This was the first time, for example, that eighth graders had met the Interpretive Panelists' expectations in satisfactorily selecting the proper word form (e.g., discover, discovery, discovering) to complete a sentence. **The scoreboard: no strengths, 1 weakness, 12 satisfactory.**

Consistent Strengths and Weaknesses Deserve Attention

Because the Alaska Statewide Assessment has measured the same basic skills for the past three assessments (1981, 1983, and 1985), and because Interpretive Panels have been used each time, it is now possible to make some year-to-year comparisons. Because neither the items nor the panelists are the same every year, the number of strengths and weaknesses changes from year to year, and skills that are strengths (or weaknesses) one year may not be in subsequent years. Indeed, the strengths which we can take the most pride in, and the weaknesses we should worry about most, are those that have appeared consistently. The accompanying table shows the skills that have been consistent strengths and weaknesses across all of the last three tests—'81, '83 and '85.

The Content Panels which review the assessment results examine the weaknesses carefully. In general, they feel that some of the weaknesses shown here are weaknesses not because instruction is poor but because the skill is not taught in all Alaska schools by the time the assessment is given. As stressed earlier in this report, there is no statewide curriculum and it is perfectly acceptable for some of the skills to not be covered by a district. But the decision not to teach a skill should be a carefully considered decision, not an accident. Educators should review these weaknesses to see if they are being taught in their district and, if they are not, whether the decision not to was intentional.

When looking at Interpretive Panel results from 1981, 1983 and 1985, one discovers some consistencies in the determination of statewide strengths and weaknesses. The following skills were identified the same way in all three years by the Interpretive Panels; 3 were consistent strengths, 23 were consistent weaknesses.

Strengths (Mathematics)

- Student will determine operation to use for solving word problem (grade 4 math application)
- Student will add numbers with decimals aligned with regrouping (grade 4 math computation)
- Student will select word form for given number (grade 4 math concepts)

Weaknesses (Mathematics)

- Student will multiply number with two digit multiplier and regrouping (grade 4 math computation)
- Student will divide numbers with one digit divisor and no remainder (grade 4 math computation)
- Student will divide numbers with two

digit divisor and remainder (grade 4 math computation)
- Student will identify odd and even numbers (grade 4 math concepts)
- Student will identify relationship between two metric or customary units (grade 4 math concepts)
- Student will determine if additional information is needed to solve a word problem (grade 8 math application)
- Student will subtract negative number from positive number (grade 8 math computation)
- Student will recognize correctness of a statement of the form $(2+A)\times4 = (2\times4)+(A\times4)$ (grade 8 math concepts)

Strengths (Reading)

(None)

Weaknesses (Reading)

- Student will use dictionary definitions to answer question (grade 4 study skills)
- Student will use dictionary guidewords to determine location of given word (grade 4 study skills)
- Student will use index to locate information (grade 4 study skills—two items)
- Student will use guide letters to determine which topic would be found in given encyclopedia volume (grade 4 study skills)
- Student will recognize best part of book to use to locate specific information (grade 4 study skills)
- Student will follow written directions to answer question (grade 4 study skills)
- Student will recognize words having same initial blend sound (grade 4 word skills)
- Student will recognize words having same initial digraph sound (grade 4 word skills)
- Student will recognize words containing diphthongs as having same vowel sound (grade 4 word skills)
- Student will recognize two consonant-controlled vowels with same sound (grade 4 word skills)
- Student will use correct regular plural noun to complete sentence (grade 4 word skills)
- Student will use correct irregular plural to complete sentence (grade 4 word skills)
- Student will determine out of date material, based on book title and publication date (grade 8 study skills)
- Student will use correct plural possessive noun to complete sentence (grade 8 word skills)

As can be seen, fourth grade reading is the area where students performed lowest according to teachers' expectations. Within that test, there were seven weaknesses in the study skills subtest and six weaknesses in the word skills subtest. In no other case were there more than three weaknesses in a subtest.

From time to time, the staff working on Alaska's Assessment Program enlist the advice and counsel of the District Testing Representatives to investigate some problem associated with testing in the state. These 55 educators–one from each district–form the link between the Assessment Program and the local staff who use the results of student testing to make sound educational decisions.

This month, the editors of Assessment Report 13 interviewed five experienced test reps:

- *Wally Berard–North Slope Schools*
- *Dave Dossett–Southeast Island Schools*
- *Chris Robinson–Southwest Region Schools*
- *Mary Rubadeau–Bering Straits Schools*
- *Ralph Rudzik–Alaska Gateway Schools*

on a topic of considerable importance–the potential for an unexpected decline in standardized test scores brought about by the recent renorming of the test series. Their perspectives, combined with some background on the problem, make up the following article.

Forewarned is Forearmed . . .
Educators May See Standardized Test Scores Drop

Sixth grade teacher Ron Johnson looks over his students' standardized test scores with disbelief. In spite of his best efforts, his class—which always scored above the national average in earlier grades—is now below the national average. "What's happening?" he asks his principal. "The students I've had have always scored better than average. Why this sudden change?"

Ernie's parents have the same question. Relative to the national average, their son's test scores are lower than in previous years. His parents had always been pleased with the instruction their son had received and had felt that his teachers were doing a good job. But now they're not so sure. Should they complain about the drop in Ernie's achievement level?

Superintendent Nelson has a bigger problem. For the last five years, the standardized test scores of his district have been rising. In fact, the district, which had previously scored in the lowest quarter of districts in the state which took the same test, was now nearing the statewide average. But the scores from this spring were a disaster, falling substantially rather than continuing their upward trend. Unfortunately, Nelson had claimed that the increasing test scores were, in part, the result of revised programs and procedures he personally implemented. "What did I do this past year that messed things up?" he wondered. "After taking the credit for the score increase, do I now have to take the blame for the score drop?"

Administrators, teachers and parents have all run into the same problem: an unanticipated—and unexplained—drop in scores relative to a comparison group. Why such a drop may occur, and how it affects educational planning, is a matter of importance to all Alaska educators who make use of standardized test scores.

(continued, next page)

Test Score Drop

(continued from page 5)

Some Background

Over the past several years, the publishers of the most widely used standardized achievement series have developed new versions of their tests. They do this regularly—every seven to ten years—in order to reflect changes in instructional practices and curricular emphases and to establish new benchmarks for student performance on the tests.

These benchmarks, called norms, are established by testing a large number of students (well over 100,000) with the new tests before the series is placed in general use. The sample used to establish the norms is very carefully chosen to be representative of the nation as a whole, several geographic regions (such as the West) and each state.

The norming is a bit like determining how an election will come out by asking a relatively small number of people how they voted. The publishers get a precise indication of how students throughout the nation will score by testing a carefully selected sample. Publishers take pains to include appropriate numbers of minority students and other special populations in their norming effort, and to weight the sample to reflect different population patterns. After all, the usefulness of these benchmarks—the norms—depends largely on how good a match there is between all students who will take the test and the smaller group which participated in the norming. Since schools spend about $100 million each year purchasing standardized tests, the market for a sound test series is big enough to warrant conducting the norming carefully.

When a district administers a standardized test, the reports come back with information about how students, schools and the district did compared to the norm group. If, for example, a student got exactly the same score as the average of the norm group, the student is said to have scored "at the fiftieth percentile." This means that half of the norm group scored better than our example student, while half scored worse. If the student's score was better than three out of four students in the norm group, the student scored at the 75th percentile.

In the past year or two, the publishers of the most widely used achievement tests—the SRA Achievement Series, the Metropolitan Achievement Tests, the Comprehensive Test of Basic Skills, the California Achievement Tests and the Iowa Tests of Basic Skills—have all developed new versions of their tests and have, as a result, "renormed" their tests. As a result, virtually every school in Alaska is using a different set of tests, and thus being compared to a different norm group, than a year or two ago.

So What Happens to the Scores?

If achievement levels have, in general, fallen in the years between the norming of the old and new versions of a test, a given achievement level will look better in comparison to the new norms. For example, a district's performance that might have been equal to the average of the old norm group might be better than 6 out of 10 scores from the new norm group. Thus, without any real change in achievement, the district has climbed from the 50th to the 60th percentile.

On the other hand, if the norm group's achievement has improved between the two normings, the same achievement level will look worse when compared to the new norms. The performance that was once average might now be better than only four of ten districts. Without any change in achievement, scores have fallen from the 50th to the 40th percentile.

In the seven or so years that have passed since most of the standardized tests were normed, many people feel that scholastic standards have risen. This seems to be particularly true in the basic skills areas (reading, mathematics, language arts, etc.) assessed by the standardized tests. More attention has been given to both the subject areas and students' performance on measures of academic achievement such as test scores. All things considered, there is a good chance that the students chosen for the publishers' most recent norm groups had a higher level of knowledge than their counterparts from seven years earlier.

In the past year or two, the publishers of the most widely used achievement tests - the SRA, the Metropolitan, the CTBS, the California Achievement Tests and the ITBS - have all developed new versions of their tests. As a result virtually every school in Alaska is using a different set of tests, and thus being compared to a different norm group than a year or two ago.

Now consider what might happen to a district over the period of years between the two normings of a test. Let's suppose that the district was slightly below the national norm group average seven years ago. Suppose that in the next year, the district's absolute performance got a little better; their relative performance, based on the old norm group, would also be better. If this process were repeated for each of the next six years, the district might find that it was doing much better relative to the norm group, now even scoring substantially above the national average.

In the next year, though, they would be using the new version of the test and, as a result, be compared to the new norm group. If that norm group has shown the same achievement increase as our example district, the district's relative score will be back where it was—below the national average. To most observers, it will look like the district's scores went from well above to below the national average in a single year. But in reality, the achievement level of the students probably didn't drop at all.

This is more than a theoretical problem. Take, for example, the plight of the Oakland, California school district, which received national attention when the dramatic drop in its test scores was documented in an Associated Press article last fall. After having its scores increase significantly in earlier years, Oakland began to use a newly-normed version of the Comprehensive Test of Basic Skills two years ago. Oakland's scores plummeted. The sixth grade dropped from the 48th to the 38th percentile, while the ninth grade dropped from the 35th to the 27th percentile. The third grade drop was even worse—from the 52nd percentile (above the national average) all the way down to the 33rd.

But Alaska Districts Are Knowledgeable, Prepared

A precipitous drop in test scores can prove very embarrassing, particularly if it comes with little advance warning or explanation. But interviews with the testing representatives of a sample of Alaska's districts show that they are aware of the potential for a drop in standardized test scores and have begun to prepare district staff. And because of this advance preparation, teachers, administrators and parents may avoid much of the shock and confusion that confronted our hypothetical score recipients at the beginning of this article.

The key, say the test reps, is to establish early on that the score decline is a possibility following national renorming of standardized achievement tests, and not a signal of potential educational problems in the district. "In general, the year of a renorming is not a good time to plan education policy based on test scores," they stressed. "Program changes could have affected the test scores, but it would be impossible to separate the impact of such changes from the effect of the renorming."

The key, say the test reps, is to establish early on that the score decline is a possibility following national renorming of standardized achievement tests, and not necessarily a signal of potential educational problems in the district.

Because some of the test series were renormed two years ago, one of the test reps had already had the experience of having to explain score declines. Fortunately the district had been alert to the possibility of score drops and was not shocked when they occurred. "Before the test was even given," commented the test rep, "teachers had been encouraged to remember that this was a different test with different norms." Even so, there were still questions from teachers and difficulty explaining the lower scores to some parents.

The four other test reps interviewed will be receiving scores based either on newly-normed editions of their test or, in one case, an entirely new test series. But the test reps and their associates were well aware that the changes will influence the apparent performance. One advised that "involving staff in all phases of the test selection is a key step toward ensuring that they would understand the anticipated differences in scores." And since testing practices in general have been emphasized in the district, not many teachers and administrators would be caught flat-footed if scores dropped.

Test publishers' representatives were of varying degrees of assistance to the test teps interviewed. In some districts, the publisher's rep hadn't mentioned anything about the possible score declines, while in others, the publisher's rep was working with the district to develop special ways of examining the data by focusing on strengths and weaknesses uncovered by the testing rather than the norm-referenced information. And while some of the test publishers have said they will provide information to allow school districts to equate scores under the old and new norms, at least one representative denies that there will be any problem, feeling that "there won't really be a diffence between the old and new norms."

There Are Ways to Reduce the Problem

How can an Alaska district prepare for the possibility that test scores will drop even though actual student achievement hasn't? The following list—compiled from the comments of district testing representatives and information contained in Alaska's recently released Assessment Handbook—gives some suggestions.

- Understand what a norm referenced test is and how it is constructed. Know why norm referenced and nationally-published standardized tests, while really two different things, are often discussed interchangeably.
- Know what terms like percentile and grade equivalent mean and use them correctly. In particular, make certain that everyone who discusses test scores with parent and community groups and the media understands these terms and can explain them correctly.
- Know where the test series you're using stands in regard to renorming. When was the renorming done and, more important, which set of scores is the first one with the new norms? In most cases it will be this spring's results, but some districts received renormed results last year.
- Get your test publisher's official word on what will happen because of the renorming. Do they expect score changes? Will they provide the information necessary to equate "old" and "new" scores? Work with your publisher's representative on this, but try to make certain he or she has given you accurate information rather than just a convenient response.

Falling test scores are going to cause the most trouble in districts where recent score gains have been highly publicized and, often, attributed to innovative programs or the skill of certain staff. In such a district, a score decline - even if it is not actually the result of poorer performance - will be hard to explain.

- Don't provide test scores in isolation to any audience; always accompany them with explanatory information. School and district scores should be examined and analyzed by a group of educators who can make sound interpretations of what the scores mean. Those interpretations—including the impact of the renorming, when appropriate—should then be made available to everyone reviewing the test scores.

Above all, remember that "if you live by the sword, you die by the sword." Metaphorically, this means that if you have equated higher norm referenced test scores with a much improved educational program, then you'll have a hard time showing why lower test scores don't mean a poorer performance. Falling test scores are going to cause the most trouble in districts where recent score gains have been highly publicized and, often, attributed to innovative programs or the skill of certain staff. In such a district, a score decline—even if it is not actually a result of poorer performance—will be hard to explain. In districts where standardized test scores are used primarily by educators for making decisions about educational programs, a fall in test scores because of a renorming can be easily explained.

Using Assessment Results in Local Districts

The interpretations presented in this newsletter are based on a statewide perspective; interpretations of students' performance district by district could differ markedly. For example, suppose that District X provides no instruction in a given skill—such as using an index, let's say. Yet 50% of the students tested in District X correctly respond to an item on using an index. Teachers and administrators in District X might view such performance in a very positive light—even if the 50% fell below the established statewide minimum set by the Interpretive Panel.

How then might individual districts look at these results? First, most will want to carefully consider the findings of the Interpretive Panel—but keep well in mind the areas of emphasis in local curriculum. In considering local versus statewide performance, it helps to ask

- Are identified strengths and weaknesses similar (that is, is the district "performance profile" about the same as that for the state as a whole)?
- Are there striking differences? (e.g., an identified weakness statewide may emerge as a strength for some districts, or vice versa.)
- How close to "cutoff" levels are strengths and weaknesses? A weakness right on the borderline—say, 1% below the established minimum level—will usually be less cause for concern than one that is, say, 20% below the minimum. (But again, local curricular emphasis is the key to identifying what's important.)

In addition, local educators will want to look closely at the minimum and acceptable performance levels set by the Interpretive Panel. What is important statewide may or may not match local priorities. In some cases, local educators may feel that minimums are too high—or too low. The point is, levels set by the Interpretive Panel are only guidelines; they are not inflexible criteria for judging all performance.

The statewide intepretive panel process can be quite useful, but local staff may also want to conduct their own interpretation process on local results. Staff can follow the same procedures for setting minimum and desired levels as are used for the statewide panels. In fact, the forms and directions necessary to conduct local interpretations are available from the Department of Education; call Al Hazelton at (907) 465-2966.

While information about strengths and weaknesses is interesting by itself, it becomes especially useful if a district has matched the assessment items to its curriculum.

While information about strengths and weaknesses is interesting by itself, it becomes especially useful if a district has matched the assessment items to its curriculum. Then it is possible to make quite specific interpretations—interpretations that can have a direct impact on curriculum planning and implementation. The Assessment Handbook recently distributed by the Department to all district offices has a chapter devoted to the role statewide assessment can play in improving local education; your District Testing Representative can provide you with a copy.

Office of Evaluation, Assessment & Research
Alaska Department of Education
Pouch F
Juneau, AK 99811

ASSESSMENT REPORT—Please Circulate To All Interested Individuals